DirectCertify

Unlock Your Potential, Certify Your Success

NVIDIA

NCA-AIIO Exam

AI Infrastructure and Operations

Exam Latest Version: 9.0

DEMO Version

Full Version Features:

• 90 Days Free Updates

• 30 Days Money Back Guarantee

• Instant Download Once Purchased

• 24 Hours Live Chat Support

Full version is available at link below with affordable price.

https://www.directcertify.com/nvidia/nca-aiio

Question 1. (Multi Select)

A large enterprise is deploying a high-performance AI infrastructure to accelerate its machine learning workflows. They are using multiple NVIDIA GPUs in a distributed environment. To optimize the workload distribution and maximize GPU utilization, which of the following tools or frameworks should be integrated into their system? (Select two)

A: NVIDIA CUDA
B: NVIDIA NGC (NVIDIA GPU Cloud)
C: TensorFlow Serving
D: NVIDIA NCCL (NVIDIA Collective Communications Library)
E: Keras

Correct Answer: A, D

Explanation:

In a distributed environment with multiple NVIDIA GPUs, optimizing workload distribution and GPU utilization requires tools that enable efficient computation and communication:

NVIDIA CUDA(A) is a foundational parallel computing platform that allows developers to harness GPU power for general-purpose computing, including machine learning. It's essential for programming GPUs and optimizing workloads in a distributed setup.

NVIDIA NCCL(D) (NVIDIA Collective Communications Library) is designed for multi-GPU and multi-node communication, providing optimized primitives (e.g., all-reduce, broadcast) for collective operations in deep learning. It ensures efficient data exchange between GPUs, maximizing utilization in distributed training.

NVIDIA NGC(B) is a hub for GPU-optimized containers and models, useful for deployment but not directly responsible for workload distribution or GPU utilization optimization.

TensorFlow Serving(C) is a framework for deploying machine learning models for inference, not for optimizing distributed training or GPU utilization during model development.

Keras(E) is a high-level API for building neural networks, but it lacks the low-level control needed for distributed workload optimization—it relies on backends like TensorFlow or CUDA.

Thus, CUDA (A) and NCCL (D) are the best choices for this scenario.

Question 2. (Single Select)

Which NVIDIA solution is specifically designed to accelerate data analytics and machine learning workloads, allowing data scientists to build and deploy models at scale using GPUs?

A: NVIDIA CUDA
B: NVIDIA JetPack
C: NVIDIA RAPIDS
D: NVIDIA DGX A100

Correct Answer: C

Explanation:

NVIDIA RAPIDS is an open-source suite of GPU-accelerated libraries specifically designed to speed up data analytics and machine learning workflows. It enables data scientists to leverage GPU parallelism to process large datasets and build machine learning models at scale, significantly reducing computation time compared to traditional CPU-based approaches. RAPIDS includes libraries like cuDF (for dataframes), cuML (for machine learning), and cuGraph (for graph analytics), which integrate seamlessly with popular frameworks like pandas, scikit-learn, and Apache Spark.

In contrast:

NVIDIA CUDA(A) is a parallel computing platform and programming model that enables GPU acceleration but is not a specific solution for data analytics or machine learning—it's a foundational technology used by tools like RAPIDS.

NVIDIA JetPack(B) is a software development kit for edge AI applications, primarily targeting NVIDIA Jetson devices for robotics and IoT, not large-scale data analytics.

NVIDIA DGX A100(D) is a hardware platform (a powerful AI system with multiple GPUs) optimized for training and inference, but it's not a software solution for data analytics workflows—it's the infrastructure that could run RAPIDS.

Thus, RAPIDS (C) is the correct answer as it directly addresses the question's focus on accelerating data analytics and machine learning workloads using GPUs.

Question 3. (Multi Select)

You have developed two different machine learning models to predict house prices based on various features like location, size, and number of bedrooms. Model A uses a linear regression approach, while Model B uses a random forest algorithm. You need to compare the performance of these models to determine which one is better for deployment. Which two statistical performance metrics would be most appropriate to compare the accuracy and reliability of these models? (Select two)

A: F1 Score
B: Learning Rate
C: Mean Absolute Error (MAE)
D: Cross-Entropy Loss
E: R-squared (Coefficient of Determination)

Correct Answer: C, E

Explanation:

For regression tasks like predicting house prices (a continuous variable), the appropriate metrics focus on accuracy and reliability of numerical predictions:

Mean Absolute Error (MAE)(C) measures the average absolute difference between predicted and actual values, providing a straightforward indicator of prediction accuracy. It's intuitive and effective for comparing regression models.

R-squared (Coefficient of Determination)(E) indicates how well the model explains the variance in the target variable (house prices). A higher R-squared (closer to 1) suggests better fit and reliability, making it ideal for comparing Model A (linear regression) and Model B (random forest).

F1 Score(A) is used for classification tasks, not regression, as it balances precision and recall.

Learning Rate(B) is a hyperparameter for training, not a performance metric.

Cross-Entropy Loss(D) is typically used for classification, not regression tasks like this.

MAE (C) and R-squared (E) are standard metrics in NVIDIA RAPIDS cuML and other ML frameworks for regression evaluation.

Question 4. (Single Select)

Your AI data center is running multiple high-performance GPU workloads, and you notice that certain servers are being underutilized while others are consistently at full capacity, leading to inefficiencies. Which of the following strategies would be most effective in balancing the workload across your AI data center?

A: Use horizontal scaling to add more servers
B: Manually reassign workloads based on current utilization
C: Implement NVIDIA GPU Operator with Kubernetes for automatic resource scheduling
D: Increase cooling capacity in the data center

Correct Answer: C

Explanation:

The NVIDIA GPU Operator with Kubernetes (C) automates resource scheduling and workload balancing across GPU clusters. It integrates GPU awareness into Kubernetes, dynamically allocating workloads to underutilized servers based on real-time utilization, priority, and resource demands. This ensures efficient use of all GPUs, reducing inefficiencies without manual intervention.

Horizontal scaling(A) adds more servers, increasing capacity but not addressing the imbalance—underutilized servers would remain inefficient.

Manual reassignment(B) is impractical for large-scale, dynamic workloads and lacks scalability.

Increasing cooling capacity(D) improves hardware reliability but doesn't balanceworkloads.

The GPU Operator's automation and integration with Kubernetes make it the most effective solution (C).

Question 5. (Single Select)

What is a key consideration when virtualizing accelerated infrastructure to support AI workloads

on a hypervisor-based environment?

A: Enable vCPU pinning to specific cores
B: Disable GPU overcommitment in the hypervisor
C: Maximize the number of VMs per physical server
D: Ensure GPU passthrough is configured correctly

Correct Answer: D

Explanation:

When virtualizing GPU-accelerated infrastructure for AI workloads,ensuring GPU passthrough is configured correctly(D) is critical. GPU passthrough allows a virtual machine (VM) to directly access a physical GPU, bypassing the hypervisor's abstraction layer. This ensures near-native performance, which is essential for AI workloads requiring high computational power, such as deep learning training or inference. Without proper passthrough, GPU performance would be severely degraded due to virtualization overhead.

vCPU pinning(A) optimizes CPU performance but doesn't address GPU access.

Disabling GPU overcommitment(B) prevents resource sharing but isn't a primary concern for AI workloads needing dedicated GPU access.

Maximizing VMs per server(C) could compromise performance by overloading resources, counter to AI workload needs.

NVIDIA documentation emphasizes GPU passthrough for virtualized AI environments (D).

DirectCertify

Unlock Your Potential, Certify Your Success

Full version is available at link below with affordable price.

https://www.directcertify.com/nvidia/nca-aiio

30% Discount Coupon Code: LimitedTime2025